Chapter 6 – Data Management

INTRODUCTION

Collecting data on specific natural resource variables is our first step toward understanding the ecosystems within our national parks. These ecosystems are changing, as is our knowledge of them and how they work. We use monitoring data to analyze, synthesize, and model aspects of ecosystems. In turn, we use our results and interpretations to make decisions about the parks' vital natural resources. Thus, *data* collected by researchers and maintained through sound data management practices will become *information* through analyses, syntheses, and modeling. Information is the common currency among the many different activities and people involved in the stewardship of NPS natural resources. Users of network generated information include park managers, cooperators, researchers, and the general public.

Data management refers to the attitudes, habits, procedures, standards, and infrastructure related to the acquisition, maintenance, and disposition of data and its resulting information. Data management is not an end unto itself, but instead is the means of maximizing the quality and utility of our natural resource information. This is particularly important for long-term programs, in which the lifespan of a data set will likely be longer than the careers of the scientists who developed it. Seen in this way, it becomes obvious that data management is vital to the success of any long-term monitoring initiative.

This chapter summarizes the system of data management that will be used by the Great Lakes Network. This system is explained more completely in the Network's Data Management Plan (DMP). See Appendix A, Supplemental Document 8 (Hart and Gafvert 2005). The complete DMP presents the overarching strategy for ensuring that program data are documented, secure, accessible, and useful for decades into the future. The plan also refers to other guidance documents, SOPs, and detailed protocols that convey specific standards and steps for achieving our data management goals. The Data Management Plan is the foundation that we will build upon as new protocols are developed, advances in technology are adopted, and new concepts in data management philosophy are accepted.

DATA MANAGEMENT GOALS AND OBJECTIVES

The goal of our data management system is to ensure the quality, interpretability, security, longevity, and availability of ecological data and related information resulting from resource inventory and monitoring efforts.

- ***Quality.*** The Network will ensure that appropriate quality assurance measures are taken during all phases of project development, data acquisition, data handling, summary and analysis, reporting, and archiving. Because standards and procedures can only accomplish so much, an important part of quality assurance is to continually encourage careful attitudes and good habits among all staff involved in creating, collecting, handling, and interpreting data.
- ***Interpretability.*** A data set is only useful if it can be readily understood and appropriately interpreted in the context of its original scope and intent. Data taken

out of context can lead to misinterpretation, misunderstanding, and bad management decisions. Sufficient documentation (e.g., metadata) will accompany each data set, and any reports and summaries derived from it, to ensure that users will have an informed appreciation of the context, applicability, and limitations of the data.

- ***Security.*** The Network will ensure that both digital and analog forms of source data are maintained and archived in an environment that provides appropriate levels of access to project managers, technicians, decision makers, and other users. Our data management system will take advantage of existing systems for Network security and systems backup, and augment these with specific measures aimed at ensuring the long-term security and integrity of our data.
- ***Longevity.*** Countless data sets have become unusable over time either because the format is outdated (e.g., punchcards) or because metadata is insufficient to determine the collection methods, scope and intent, quality assurance procedures, or format of the data. Proper storage conditions, backups, and migration of data sets to current platforms and software standards are basic components of data longevity. Comprehensive data documentation is an essential component of data management. The GLKN will use a suite of metadata tools to ensure that data sets are consistently documented, and in formats that conform to current federal standards.
- ***Availability.*** Natural resource information can only inform decisions if it is available to managers at the right time and in a usable form. Our objective is to expand the availability of natural resource information by ensuring that the products of inventory and monitoring efforts are created, documented and maintained in a manner that is transparent to the potential users of these products. The Network will endeavor to provide natural resource managers easy, secure, and continuous access to its data and analyses, based on the users' needs.

DATA MANAGEMENT ROLES AND RESPONSIBILITIES

For the GLKN monitoring program to work effectively, all employees will have data stewardship responsibilities. The GLKN Data Management Plan specifies the roles and responsibilities of individuals involved in the production, analysis, management, and reporting of data and information. This includes field workers, natural resource specialists, program ecologists, GIS specialists, and other specialists such as biometricians. More detailed roles and responsibilities are given in the protocol for each Vital Sign. Table 6.1 lists the basic roles and responsibilities of individuals involved in monitoring projects, although not all of them will be involved in every project and individuals may assume multiple roles. For example, a network ecologist may have a role in developing a project protocol and ongoing involvement in issues related to ecological science, and at the same time may serve as the project manager for the same project.

Chief personnel involved with data management include the project manager and the data manager. Figure 6.1 illustrates the core data management duties of the project manager and data manager and where they overlap.

Table 6.1. Roles and responsibilities for data stewardship.

Role	Data Stewardship Responsibilities
Project Manager	Oversee and direct project operations, including data management
Project Crew Leader	Supervise crew members and organize data
Project Crew Member	Collect, record, and verify data
Network Ecologist	Integrate science with Network data and activities
Network Coordinator	Coordinate and supervise all Network activities
Network Data Manager	Ensure inventory and monitoring data are organized, useful, compliant, safe, and available
Database Specialist	Know and use database software and database applications
Network GIS Manager	Support Network objectives with GIS and resource information
GIS/Data Specialist	Process and manage data
Information Technologist	Provide IT support for hardware, software, and networking
Statistician or Biometrician	Analyze data and/or consult on analysis
Park Research Coordinator	Facilitate research and data acquisition in a park. Communicate NPS and park requirements to permit holders
Curator	Oversee all aspects of specimen acquisition, documentation, preservation, and use of park collections
I&M National Data Manager	Provide Servicewide database availability and support
End Users (managers, scientists, interpreters, and public	Inform the scope and direction of science information needs and activities. Apply data and information services and products.

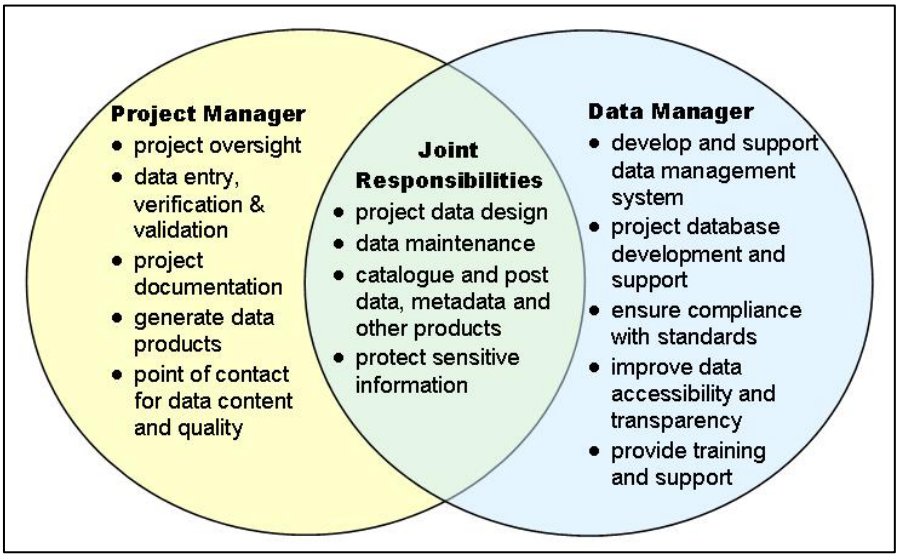

Figure 6.1. Core data stewardship duties of project managers and data managers.

PROJECT WORK FLOW AND THE DATA MANAGEMENT PROCESS

Both short-term and long-term projects share many work flow and data management characteristics. Most GLKN projects consist of five primary stages: planning and approval; design and testing; implementation; product integration; and dissemination of information, evaluation, and closure. Each stage is characterized by a particular set of activities.

- ***Planning and Approval.*** Establishing the project scope and objectives is the most important step in project development. It is crucial that Network and park staff work together at this stage to establish what data are needed, why they are needed, how they will be used, and any unique data management requirements.
- ***Design and Testing.*** At this stage, specifications are established for how data will be acquired, processed, analyzed, reported, and made available to others. The project manager and data manager work together to develop specific procedures (SOPs) related to data acquisition, processing, analysis, and quality control. Also, the project manager and data manager collaborate to develop the data design and data dictionary, in which the specific variables that will be collected are described in detail. In addition, decisions should be made regarding integration and permanent storage of deliverables as they are produced.
- ***Implementation***. During the implementation phase, data are acquired, processed, error-checked and documented. Although data collection and processing methods will vary among projects, each project will require data verification and validation. All aspects of data acquisition should be specified in project protocols and SOPs. Similarly, metadata should include documentation of quality assurance measures. During this phase, the data are preliminary and available only to individuals involved in the project.
- ***Product Integration and Data Dissemination.*** During this stage, data products and other deliverables are integrated into national and Network databases, metadata records are finalized and posted to clearinghouses, and products are distributed or made available to the project's intended audience(s). This is also when items that belong in collections or archives are accessioned and cataloged. Certain projects, such as those conducted jointly with other agencies and using a common database, may have additional integration needs.
- ***Evaluation and Closure.*** For long-term monitoring and other cyclic projects, this phase occurs at the end of each field season and leads to an annual review of the project. After products are cataloged and made available, program administrators, project managers, and data managers will assess how well the project met its objectives, determine what might be done to improve various aspects of the project methodology, and evaluate the usefulness of the resulting information.

Following evaluation, changes will be incorporated into the protocol as needed. This may necessitate redesign and testing or simply a procedural change in the implementation phase. The evaluation process involves feedbacks and reassessments, in essence becoming an iterative process.

STRATEGIES FOR DATABASE DESIGN

Long-term monitoring projects conducted by the Network will have modular, stand-alone project databases that share design standards and centralized validation tables. The project databases will be developed in a desktop database application or a GIS geodatabase application, depending on the requirements of a project and the desires of the project manager. Because all natural resource monitoring consists of observations and measurements taken at specific geographic locations, nearly all the Network's monitoring data sets are inherently suited to management in a geographic information system (GIS). The Network's data management vision involves maintaining a close spatial link to associated monitoring data in a format that allows it to be readily visualized in a geographic context. A generic name for a spatially explicit data structure is a geodatabase. There are numerous advantages to maintaining project-specific (geo)databases:

- Data sets are modular, allowing greater flexibility in accommodating the needs of each project area. By having project-specific data sets, databases and protocols can be developed at different rates without a significant cost to data integration. In addition, one project database can be modified without affecting the functionality of other project databases.
- By working up from modular data sets, we avoid a large initial investment in a centralized database and the concomitant difficulties of integrating among project areas with very different – and often unforeseen – structural requirements. Furthermore, the initial investment in integration may not result in greater efficiency in the future.

Standards for project databases ensure compatibility among data sets, and are essential given the often unpredictable ways in which data sets are aggregated and summarized. When well conceived, standards encourage sound database design and facilitate interpretability of data sets. Shared 'lookup' tables (e.g., species lists, park names, common location information) help standardize data and facilitate integration. As much as possible, GLKN standards for fields, tables, and other database objects will mirror those conveyed through the Natural Resource Database Template. Where differences between local and national standards exist, the rationale for these differences will be documented. In addition, documentation and database tools (e.g., queries that rename or reformat data) will be developed to ensure that data exports for integration are in a format compatible with current national standards.

Although stand-alone databases work well at the project level, they are not efficient for analysis across ecological indicators and at the ecosystem or park management level. In addition, the Network must make its data sets readily available to appropriate user groups. Because GIS software vendors have focused on making data transfer between desktop geodatabases and enterprise geodatabases very efficient, the opportunity to combine the project databases into a SQL (Structured Query Language) database that contains all the common and unique tables of each project offers promise for reducing the Network's data management tasks and giving the Network's data users a single access point. The Network is developing a web portal as its primary mechanism

for distributing data. The site is based on an enterprise SQL relational database and includes an Internet Mapping Service (IMS).

DATA AND INFORMATION INFRASTRUCTURE

The GLKN program relies in part on park, regional, and national IT personnel and resources to maintain the computer resource infrastructure. This includes, but is not limited to, hardware replacement, software installation and support, security updates, virus-protection, telecommunications networking, and backups of servers. Therefore, communication with park and regional IT specialists is essential to ensure service continuity for our system architecture.

An important element of a data management program is a reliable, secure network of computers and servers. Our digital infrastructure has three main components: a network-based local area network (LAN), network data servers, and servers maintained at the national level (Figure 6.2). This infrastructure is maintained by Network and national IT specialists, who administer all aspects of system security and backups.

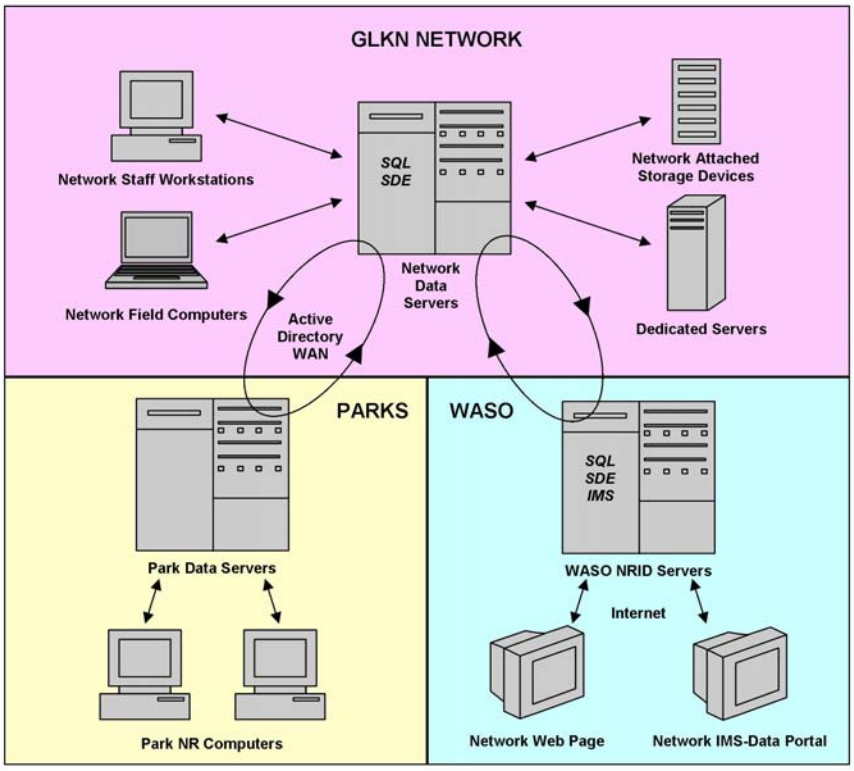

Figure 6.2. Schematic representing the logical layout and connectivity of computer resources within the GLKN core user groups. Each of these components hosts different parts of our natural resource information system.

National-level Infrastructure

Data management support from the Washington office includes hosting and maintaining several databases for summarizing park natural resource data at the national level. These online applications include:

- *NatureBib* – the master database for natural resource bibliographic references

- *Biodiversity Data Store* - a digital repository of documents, GIS maps, and data sets that contribute to the knowledge of biodiversity in National Park units, including presence/absence, distribution, and abundance
- *NPSpecies* – a biodiversity database application that lists the species that occur in or near each park, and the physical or written evidence for the occurrence of the species (i.e., references, vouchers, and/or observations)
- *NR-GIS Data Store* – a centralized repository and graphical search interface that links data set metadata to a searchable data server on which data sets are organized by NPS units, offices and programs

Region-level Infrastructure

The Midwest Region contributes to the inventory and monitoring infrastructure through higher-level networking and communications support, and the participation of the regional GIS coordinator and associated staff. The Network has GIS expertise in Network staff positions; however, several Network parks lack GIS personnel. The Regional GIS Support Office has been cooperating with GLKN to assist member parks with training, GIS project needs, and remote sensing projects. In addition, the regional GIS staff provides the Network with the first level of technical support for some GIS software applications.

Network-level Infrastructure

The Network has implemented a central server system to provide access to shared information resources. The strategy is to maintain a relational database management system (RDBMS) that allows for central management of common tables and high-value, long-term project databases and provides a means of maximizing performance in a distributed, multi-user environment. This is part of the Network's enterprise GIS/SQL strategy. The following types of materials are maintained on these Network data servers:

- Enterprise SQL Database – this will be used primarily to feed the Network's web portal, which will be used both internally and externally, but it will also provide a single source for the compiled data sets for monitoring projects and other multi-year efforts that have been certified for data quality
- Common lookup tables and data sets – for example, parks, projects, personnel, species, base GIS resources
- Project tracking application – used to track the status, deliverables, due dates, and responsibilities for each monitoring protocol
- Network digital library – Network repository for finished versions of project deliverables for Network projects (e.g., reports, methods documentation, data files, metadata, etc.)

Data redundancy, use of data servers, and distribution of final products are highlights of GLKN's information management infrastructure. Redundancy means that data are fully backed up and stored at an off-site location. This is crucial for information recovery in case of a local catastrophe at one of the host sites. Backups will be automated through scheduled services. Data servers will act as a repository for data and data

products generated by the program. These data will be accessible to authorized personnel via an IMS web portal being developed by Michigan State University and Colorado State University. Security permissions will be granted down to the project level and access to preliminary or sensitive data will be carefully controlled. Finalized data products and related information will be uploaded to online national databases (NatureBib, NPSpecies, NR-GIS Metadata Database, and NR-GIS Data Store) for public access.

Given our collaboration with other agencies and organizations, certain GLKN data sets may be maintained by outside organizations. In such cases, we will maintain local copies of metadata for these data sets. In cases where access to the information systems supported by cooperators do not meet the Network's needs, versioned copies of data sets may be maintained on our servers to ensure data availability.

Park-level Infrastructure

Because GLKN work is largely conducted in the member parks, for the primary purposes of informing resource managers, the Network has a high degree of data exchange with its parks. Information resources shared between the Network and parks include:

- Local applications – desktop versions of database applications for a specific Network or park need
- Working files – draft geospatial themes, drafts of reports, administrative records
- Park digital library – base spatial data, imagery, and finished versions of park project deliverables
- Park GIS files – base spatial data, imagery, and project-specific themes

DATA LIFE CYCLE

The types of data handled by the Network fall into three general classifications:

- Program data are produced by projects that are either initiated (funded) by the I&M Program or involve the I&M Program in another manner (e.g., natural resource inventories and Vital Signs monitoring projects).
- Non-program legacy/existing data are produced by NPS entities without the involvement of the I&M Program (e.g., park inventory projects).
- Non-program external data are produced by agencies or institutions other than the National Park Service (e.g., weather and some water quality data).

The life cycle of data sets from each of these sources could vary considerably. For instance, partner climatic data may be acquired with considerable quality assurance and quality control review and with complete data documentation, negating the need for the Network to duplicate these tasks.

Data Acquisition and Processing

Past investments in natural resource data collection in the GLKN parks have resulted in a legacy of products that vary widely in format, consistency, and value for park stewardship. The Network has invested substantially in identifying and documenting

these legacy data sets, and in cases where they were of potential future benefit to monitoring, efforts have been made to bring data sets into compliance with current Network and NPS data standards. To help address the volume of natural resource data stored at the parks, the Network currently supports activities to obtain, catalog, report, and archive data in NPSpecies, NatureBib, and in metadata catalogs. Future work and expense to link legacy data with management requirements will be carefully scrutinized by the Network and park natural resources staffs to evaluate its potential value to current and future projects and management. Although initial GLKN-funded inventories have been completed, the Network and its member parks will continue to perform inventories according to the spirit and goals of the Natural Resource Challenge when funding is available.

In order to provide a synthesis of scientific information based on Vital Signs and related data, the Network also gathers and processes relevant data and information from other park-based and external inventory and monitoring efforts. In some cases, access to these external data sources may require the Network to enter into agreements or memoranda of understanding, or purchase subscriptions.

Most data acquired by the Network will be collected as field data (inventories and monitoring studies). Tools and methods for field data collection, such as paper data forms, field computers, automated data loggers, and GPS units will be specified in individual monitoring protocols and study plans. Various factors will determine what methods and tools are used in the field, including: data quality, security, efficiency, and a project manager's comfort level with the method employed. Field crew members will closely follow the established SOPs in the project protocol.

Quality Assurance

Long-term monitoring is only useful if users have confidence in the data. Efforts to detect trends and patterns in ecosystem processes require high-quality, well-documented data that minimize error and bias. Data of inconsistent or poor quality can result in loss of sensitivity and lead to incorrect interpretations and conclusions.

NPS Director's Order #11B: Ensuring Quality of Information Disseminated by the National Park Service (www.nps.gov/policy/DOrders/11B-final.htm) specifies that information produced by the NPS must be of the highest quality and be based on reliable data sources that are accurate, timely, and representative of the most current information available. Therefore, GLKN will establish and document procedures for quality assurance (QA) and quality control (QC) to identify and reduce the frequency and significance of errors at all stages in the data life cycle. Under these procedures, the progression from raw data to verified data to validated data implies increasing confidence in the quality of those data. Quality assurance and quality control procedures will document internal and external review processes and include guidance for handling problems with data quality.

Although the specific QA/QC procedures employed will depend on the Vital Signs being monitored, some general concepts apply to all Network projects. Examples of QA/QC practices include:

- Standardized field data collection forms
- Use of field computers and automated data loggers

- Proper calibration and maintenance of equipment
- Field crew and data technician training
- Database features such as built-in pick lists and range limits to reduce data entry errors
- Automated error-checking routines

We appraise data quality by applying verification and validation procedures. Data verification checks that the digitized data match the source data, and data validation checks that the data make sense. The Data Management Plan describes several methods for verifying and validating data, and each monitoring protocol will include specific procedures for assuring data quality.

A final report on data quality will be incorporated into the documentation for each project. Such documentation will include a listing of the specific methods used to assess data quality and an assessment of overall data quality prepared by the project manager.

Data Documentation

Data documentation is a critical step toward ensuring that all data sets retain their integrity and utility well into the future. Data documentation refers to the development of metadata. At the most basic level, metadata is 'data about data'. More specifically, it is information about the content, context, structure, quality, and other characteristics of a data set. Without meaningful metadata, potential users of a data set have little or no information regarding the quality, completeness, or manipulations performed on a particular 'copy' of a data set. Additionally, standardized metadata provide a means to catalog data sets within intranet and internet systems, thus making them available to a broad range of potential users.

At a minimum, GLKN will require the following elements for documentation of all data managed by the Network:

- Formal metadata compliant with Federal Geographic Data Committee (FGDC) standards, the National Biological Information Infrastructure (NBII) Profile (where appropriate), and the NPS Metadata Profile for all geospatial and biological data sets
- Project documentation, including data dictionaries.

The Network will create all metadata according to NPS standards and guidelines. Formal metadata will be created using ArcCatalog in conjunction with NPS Metadata Tools and Editor. The Network will publish all of its metadata to the online NR-GIS Metadata Data Store. All documentation will also be maintained with its accompanying data set(s) on the Network's data server and its web portal for data visualization and dissemination.

Data Dissemination

One of the most important goals of the I&M Program is to integrate natural resource inventory and monitoring information into NPS planning, management, and decision making. To that end, the Network will use a variety of data and information systems and employ tools that allow potential users to browse, query, and obtain data,

information, and supporting documents easily. The primary system that the Network will use for data access is its IMS web portal. In addition to on-screen visualization, the IMS website allows data sets and their metadata to be downloaded based on custom queries in industry standard file formats (e.g., MS Excel and delimited text file structures). Other data access systems include the GLKN's data server and digital library, the Network's website, and national applications with internet interfaces (NatureBib, NPSpecies, NR-GIS Data Store, etc.)

Network products will be available on request and will be distributed using file transfer protocol (FTP), attaching reports and other products with small file sizes to email, and shipping digital media such as DVDs, CD-ROMs and other disposable data storage products.

Data Maintenance, Storage, and Archiving

The Network will implement procedures to protect information over time. These procedures will ensure that digital and analog data and information are:

- up-to-date in content and format so they remain easily accessible and usable, and
- protected from catastrophic events (e.g., fire and flood), user error, hardware failure, software failure or corruption, security breaches, and vandalism.

Technological obsolescence is a significant cause of information loss, and data can quickly become inaccessible to users if they are stored in out-of-date software programs, on outmoded media, or on deteriorating (aging) media. Effective maintenance of digital files depends on the proper management of a continuously changing infrastructure of hardware, software, file formats, and storage media. As software and hardware evolve, data sets must be consistently migrated to new platforms or saved in formats that are independent of specific software or platforms (e.g., ASCII delimited text files). Storage media should be refreshed (i.e., copied to new media) on a regular basis, depending upon the life expectancy of the media.

Regular backups of data and off-site storage of backups are the most important safeguards against data loss; therefore, the Network has established data maintenance and backup schedules for data stored on the Network data servers. Although each staff member is required to backup data on personal workstations, active computers connected to the Network LAN are included in a scheduled backup one night each week.

WATER QUALITY DATA

Water quality data, including macroinvertebrate characteristics, are managed according to guidelines from the NPS Water Resources Division (Figure 6.3). These guidelines include using the NPSTORET desktop database application at the parks to help manage data entry, documentation, and transfer. The Network oversees the use of NPSTORET according to the Network's water quality monitoring protocols and ensures the content is transferred at least annually to NPS Water Resource Division for upload to the EPA STORET (STOrage and RETrieval) database.

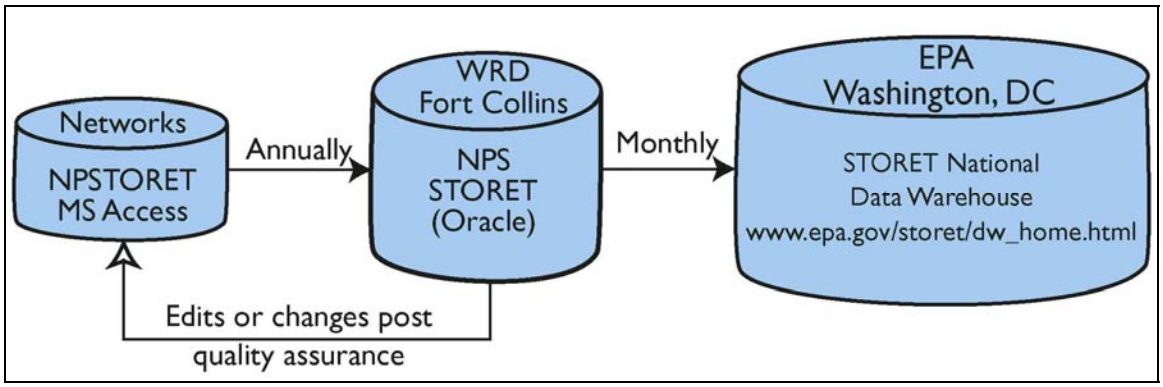

Figure 6.3. Water Quality Flow Diagram.

DATA OWNERSHIP AND SENSITIVITY

Network data and information products are property of the NPS. However, the Freedom of Information Act (FOIA) establishes the right for any person to access federal agency records that are not protected from disclosure by any exemption or by special law enforcement record exclusions. The GLKN complies with all FOIA strictures regarding sensitive data. A number of laws and regulations (see NPS Director's Order #66) allow for restricted access to information the may imperil a resource if released. Through these regulations, information that could result in harm to natural resources can be classified as 'protected' or 'sensitive' and withheld from public release (National Parks Omnibus Management Act (NPOMA).

Project managers, in conjunction with the appropriate park staff, determine data sensitivity in light of federal law and stipulate conditions for release of the data in the project protocol and metadata. The investigators, whether Network staff or partners, will develop procedures to flag information related to sensitive resources in all products, including documents, maps, databases, and metadata.